Research Memorandum 76-24

EMPIRICAL COMPARISON OF CRITERION REFERENCED MEASUREMENT MODELS

Kenneth I. Epstein and Frederick H. Steinheiser, Jr.

UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

U. S. Army

Research Institute for the Behavioral and Social Sciences

October 1976

DISPOSITION FORM

For use of this form, see AR 340-15, the proponent agency is TAGCEN.

REFERENCE OR OFFICE SYMBOL	SUBJECT
PERI-TP	Clearance and Transmittal of Reports to DTIC

TO DDC-DAA-1
ATTN: Mr. Schrecengost

FROM ARI Rsch Pub Group

DATE 8 Nov 79

CMT 1

Ms Price/48913

1. The reports listed on Inclosure 1 are approved for public release with unlimited distribution (50 numbered ARI Research Memorandums, 74-1 thru 76-30).

2. These are among the previously unrecorded ARI reports which you identified to us 22 June 1979 as not in your retrieval system. The accompanying box contains at least one copy of each report for your retention and reproduction.

Helen S. Price

HELEN S. PRICE
Research Publications Group
Army Research Institute

1 incl
List of reports, 1974-76

A

OBSOLETE.

Army Project Number
16 2Q762717A764

Unit Standards and
Performance Evaluation

9 Research ~~Memorandum~~ memo 76-24

6 EMPIRICAL COMPARISON OF CRITERION REFERENCED MEASUREMENT MODELS

10 Kenneth I. Epstein [redacted] Frederick H. Steinheiser, Jr.

Angelo Mirabella, Work Unit Leader

12 122

Submitted by:
F. Harris, Chief
UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

14 ARI-RM-76-24

11 Oct[redacted]76

Approved by:

Joseph Zeidner, Director
Organizations and Systems Research
Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for
the Behavioral and Social Sciences

EMPIRICAL COMPARISON OF CRITERION REFERENCED MEASUREMENT MODELS[1]

ABSTRACT. The Army needs information about how well an individual can perform the tasks necessary for him to do his job. This information is often gathered by means of a "criterion-referenced test," a test made up of items directly related to the job of interest. The test results can be used in two ways. The first way is to sort individuals into two groups, one made up of those who can perform their job satisfactorily and the other made up of those who do not meet minimal job requirements. A second use of the test results is to estimate the "true" capability of the examinees to do the task being tested. These two uses are clearly related. If one can precisely estimate an individual's capability, then forming the two groups is not a problem. On the other hand, it may be possible to effectively form the two groups without getting good estimates of "true" capability.

Several psychometric models are available for grouping the individuals and/or for estimating "true" scores. For example, one may simply calculate the proportion of items correctly answered and use that proportion as an estimate of "true" capability. Alternatively, a binominal error model for deriving the expression for the regression of "true" score on observed score can be used and a "true" score calculated for each individual. Other possible models include a Bayesian Model II approach and a latent trait model such as the Rasch one parameter logistic model. Each of these models yields a somewhat different estimate of "true" capability for any given individual. It follows that the makeup of the job ability groups will vary from model to model. The purpose of this research is to empirically study the models referred to above. What is needed is an appropriate statistic (or statistics) and research design for comparing each model against all others given the same test data.

I. INTRODUCTION. The purpose of this paper is to elaborate on some technical details and to highlight specific statistical and research problems introduced in a previous paper by one of the authors (Epstein, 1975).

Epstein described four procedures for estimating true scores from observed scores. The first uses the observed proportion correct as an estimate of the true proportion correct. This procedure is straightforward and familiar. Hence, discussion of it will be reserved until

---

[1] Reprinted from the Proceedings of the Twenty-First Conference on the Design of Experiments in Army Research Development and Testing, sponsored by The Army Mathematics Steering Committee on behalf of the Chief of Research, Development and Acquisition, 22-24 October 1975.

the problem of comparing the models is developed. The other three procedures are 1) a binomial error model, 2) a Bayesian model, and 3) the Rasch logistic model. Each will be discussed in detail.

2. BINOMIAL ERROR MODEL. The binomial error model (Lord and Novick, 1968, pp. 508-529) is based on the assumption that the conditional distribution of observed score for given proportion correct true score (T) is the binomial distribution.

$$h(x|T) = \binom{n}{x} T^x (1-T)^{n-x}$$

x=0,1...n is the number of correct responses observed and n is the total number of items on the test.

It is assumed that items are scored dichotomously, that total score for an examinee is the number of items answered correctly, that items are locally independent, and that items are equally diffitult for a given examinee.

The relationship between the observed score distribution and the underlying true score distribution can be written as follows:

$\phi(x) = \binom{n}{x} \int_0^1 g(T)\, T^x (1-T)^{n-x}\, dT$, x=0,1,...n, where $\phi(x)$ is the distribution of observed scores and g(T) is the unknown distribution of true scores.

It can be shown that if the regression of true score on observed score is linear then the distribution of observed score, symbolized h(x) to distinguish this special case from the general case $\phi(x)$, is negative hypergeometric.

$$h(x) = \frac{b^{[n]}}{(a+b)^{[n]}} \frac{(-n)_x}{(-b)_x} \frac{(a)_x}{x!} \qquad x = 0,1...n,$$

where

a and b are parameters to be determined and

$n^{[x]} = n(n-1)...(n-x+1)$,

$(a)_x = a(a+1)...(a+x-1)$, $n^{[0]} = (a)_0 = 1$.

The parameters, a and b, can be expressed in terms of moments of the observed score distribution

$$a = (-1+1/\alpha_{21})\, \mu_x$$

$$b = -a-1+n/\alpha_{21}$$

$$\alpha_{21} = \frac{n}{n-1} \left[1 - \frac{\mu_x(n-\mu_x)}{n\, \sigma_x^2}\right]$$

The discussion thus far has outlined an internal check of the appropriateness of this model for any given data set. That is, if one can show adequate fit to the negative hypergeometric distribution by the observed scores then it is reasonable to continue with this model assuming linear regression. If adequate fit is not obtained then either the more general nonlinear regression approach must be used or alternative models must be identified.

It can be shown that if the observed score distribution is negative hypergeometric, the true score distribution is either the two parameter beta distribution, or some other distribution having identical moments up through order n. In either case, the regression of true score on observed score is given by the linear equation

$$E(T|x) = \frac{\alpha_{21} x}{n} + \frac{(1-\alpha_{21})\mu_x}{n}, \quad x = 0,1,\ldots n.$$

3. BAYESIAN MODEL. The Bayesian model used to evaluate these data is described by Lewis, Wang, and Novick (1973). The procedure transforms the binomial test score data via an arc sine transformation. The resulting score is assumed to be a sample from a normal population with its mean value at the individual's transformed true ability. Distributions for the prior mean and variance of the examinee group's transformed scores are specified and posterior values calculated. Finally, the posterior marginal distributions for the transformed scores are obtained and estimates of individual true abilities on the original (proportion correct) scale are calculated. The mathematical details are outlined below.

The Freeman-Tukey transformation for binomial data is used in this procedure:

$$g_j = \frac{1}{2} \sin^{-1}\sqrt{\frac{x_j}{n+1}} + \sin^{-1}\sqrt{\frac{x_{j+1}}{n+1}}, \quad x_j = 1,2,\ldots n = \text{the}$$

number of correct responses. The g_j are assumed to be normally distributed with mean $\gamma_j = \sin^{-1}\sqrt{\pi_j}$ and variance $v = (4n+2)^{-1}$, where γ_j is the transformed value of the true proportion of correct responses, π_j. The validity of the assumption of normality and the suitability of the transformation for the procedures to follow can be shown to be adequate for examinee groups of at least 15 persons and for tests at least 8 items long.

The set of transformed variables, γ_j, is assumed to be a random sample from a normal distribution with mean μ_Γ and variance $\phi_\Gamma \cdot \mu_\Gamma$ and ϕ_Γ are further assumed to be independent and to have a uniform and inverse chi-square distribution respectively. Explicit expressions for the prior and posterior density functions are given in the Lewis, et al. paper.

The desired result of an analysis of this kind is the marginal posterior density function for γ_j . Unfortunately, an explicit expression for it is not obtainable from the joint posterior probability density function of the γ_j vector given the g_j vector. Lewis et al. show methods for obtaining the marginal means and variances for the γ_j using numerical integration. However, they indicate that for large sample sizes, the conditional posterior distribution of γ_j given ϕ_Γ and the g_j vector provides an acceptable approximation. The conditional approximation was used for the analysis of the data reported in the Epstein paper.

The conditional distribution of γ_j given ϕ_Γ and the g_j vector can be shown to be normal with mean

$$E\left(\gamma_j \middle| \tilde{\phi}_\Gamma , \underset{\sim}{g}\right) = \frac{\tilde{\phi}_\Gamma\, g_j + v g.}{\tilde{\phi}_\Gamma + v} ,$$

and variance

$$\operatorname{var}\left(\gamma_j \middle| \tilde{\phi}_\Gamma , \underset{\sim}{g}\right) = \frac{v(\tilde{\phi}_\Gamma + m^{-1} v)}{\tilde{\phi}_\Gamma + v} ,$$

where

$j = 1,2 \ldots m$ = the number of examinees,

$\underset{\sim}{g}$ = the vector of transformed scores, and

$\tilde{\phi}_\Gamma$ = the mode of ϕ_Γ given $\underset{\sim}{g}$.

$\tilde{\phi}_\Gamma$ can be obtained by solving the following equation:

$$(m + v + 1)\, \tilde{\phi}_\Gamma^{\,3} + \left[(m + 2v + 3)\, v - \sum_i (g_j - g.)^2 - \lambda\right] \tilde{\phi}_\Gamma^{\,2}$$

$$+ \left[(v + 2)\, v^2 - 2 \lambda v\right] \tilde{\phi}_\Gamma - \lambda v^2 = 0 .$$

In the above equation, v is the degrees of freedom for the prior inverse chi-square distribution of ϕ_Γ . Lewis, et al. recommend that a value of eight be used for most practical applications. λ is the scale factor for the inverse chi-square distribution. It can be calculated by using the formula

$$\lambda = \frac{v - 2}{4(t+1)}$$

where t is interpreted as the number of test items that the prior information is considered to be equivalent to.

Once the γ_j have been calculated, the last step in the procedure is to calculate the estimates for the true proportion correct. This is accomplished by applying the following equation:

$$\pi_j = (1 + \frac{1}{2n}) \sin^2\gamma_j - \frac{1}{4n}$$

4. RASCH MODEL. The Rasch one parameter logistic model (Wright and Panchapakesan, 1969) assumes that the observed response a_{ni} of person n to item i is governed by a binomial probability function of person ability Z_n and item easiness E_i. The probability of a correct response is:

$$P(a_{ni} = 1) = \frac{Z_n E_i}{1+Z_n E_i} .$$

The probability of a wrong response is:

$$P(a_{ni} = 0) = 1 - P(a_{ni} = 1) = \frac{1}{1+Z_n E_i} .$$

These equations may be combined to yield

$$P(a_{ni}) = \frac{(Z_n E_i)^{a_{ni}}}{1+Z_n E_i} .$$

If we let $b_n = \log Z_n$ and $d_i = \log E_i$,

then

$$P(a_{ni}) = \frac{\exp(a_{ni}(b_n + d_i))}{1 + \exp(b_n + d_i)}$$

The number of correct responses to a given set of items is the only information needed to estimate person ability. All persons who get the same score will be estimated to have the same ability. Hence, in terms of score groups,

$$P(a_{ni}) = \frac{\exp(a_{ni}(b_j + d_i))}{1 + \exp(b_j + d_i)}$$

where j = score of person n, and all persons with a score j are estimated to have the same probability governing their responses to item i.

The equations obtained when the condition of a maximum likelihood is satisfied for the model described in the preceding equation are:

$$a_{+i} = \sum_{j}^{k-1} (r_j \exp(b_j^* + d_i^*)/(1+\exp(b_j^*+d_i^*))), \quad i = 1,2,\ldots k$$

$$j = \sum_{i}^{k} (\exp(b_j^* + d_i^*)/(1+\exp(b_j^* + d_i^*))), \quad j = 1,2,\ldots k-1$$

where a_{+i} = number of persons who get item i correct

j = the total test score, an ability estimate is obtained for each score

r_j = number of persons in score group j.

b_j^*, d_i^* = estimates of b_j and d_i

The method consists of computing d_i^* and b_j^* from the implicit equations above. The equations are handled as two independent sets and solved accordingly.

An approximation of a standard error for item estimates can be obtained by assuming that the variance of the item estimate is due primarily to the uncertainty in the item score a_{+i}. To a first approximation this gives:

$$V(d_i^*) \sim (\partial d_i/\partial a_{+i})^2 \, V(a_{+i})$$

which leads to:

$$V(d_i^*) = 1/\sum_{j}^{k-1} (r_j \exp(b_j^*+d_i^*)/(1+\exp(b_j^* + d_i^*))^2).$$

The major contribution to the error variance of the ability estimate comes from the variance in scores produced by a given individual. This part of the error variance depends upon the number of items and their easiness range.

An approximation of the variance of the ability estimate b^* is given by

$$V^*(b^*) = \{1/C(b^*)\exp(b^*)\} + \{1/C^2(b^*)\}$$
$$\cdot \sum_{i}^{k} (V(d_i)\{\exp(d_i)/(1+\exp(d_i+b^*))^2\}^2)$$

where $C(b^*) = \sum_{i}^{k} (\exp(d_i)/(1+\exp(b^*+d_i))^2)$,

$V(d_i)$ is the variance of the item calibration d_i.

The first term in the denominator of the V*(b*) equation is due to the variance in the score, and the second term is due to the imprecision of item calibration. The first term is always larger than the second.

5. DISCUSSION OF THE PROBLEM. One characteristic of a useful model is that it has a small error of measurement. That is, the distribution of estimated scores for a given true score is closely clustered around the true score. The extent of the measurement error that can be expected with a given model is dependent on the variance of the estimated true score. For example, in the proportion correct model, the variance of the estimated true proportion correct is equal to $p(1-p)/n$. In this case the variance of the estimate will decrease as the number of observations increases. Thus it would seem that any level of precision could be obtained by simply adding observations. Unfortunately, for the number of items that are usually practical on a test, the level of precision possible is not completely satisfactory. It would be useful to compare the variance of the true score estimates obtained with the other models to the proportion correct model.

Therefore the question of how to derive an expression for the variance of the estimated true scores for the other models must be addressed. An expression for the binomial error model has been derived. Since the binomial error model results in a regression equation it seems reasonable to base the derivation on the general form of the error of estimation, $\sigma_E^2 = \sigma_T^2 \sqrt{1 - \rho_{xT}^2}$. The ratio of the variance of true scores to the variance of observed scores equals the reliability coefficient, $\frac{\sigma_c^2}{\sigma_x^2} = \alpha_{21}$, where σ_c is the variance of the true number correct. Since the true number correct equals the true proportion correct times the number of items, $C = nT$, one may write $\sigma_c^2 = n^2 \sigma_T^2$. Substituting, $\sigma_T^2 = \sigma_x^2 \alpha_{21}/n^2$. The reliability of a test equals the square of the correlation between true and observed scores, $\alpha_{21} = \rho_{xT}^2$. Hence, the variance of the estimated true score can be written

$$\sigma_E^2 = \frac{\sigma_x^2 \alpha_{21} (1 - \alpha_{21})}{n^2}$$

For the Bayesian and Rasch models expressions for the variances of the estimated true scores were not derived. In the case of the Bayesian model the output is in terms of the arc sine of the true proportion correct. While the sampling distribution of the transformed variable is known, the variance of the estimated true proportion correct itself was not determined. A similar problem exists for the Rasch model. The sampling distributions of the ability and item difficulty indices

are known as well as the explicit equation for calculating the proportion correct from those values. But an expression for the estimated true proportion correct has not been derived. In short, the problems are:
(1) For the Bayesian model, given the variance of a_j and the equation

$\Pi_j = (1 + 1/2n) \sin^2 \gamma_j - 1/4n$, what is the variance of Π_j ; and

(2) For the Rasch model given the variances of b* and d* and the equation $p\ (\text{correct}) = \frac{\exp(b^* + d^*)}{1 + \exp\ (b^* + d^*)}$, what is the variance of p?

As a result of the discussion during the session a solution to the above mathematical problems seems to be available. It was pointed out that methods exist for deriving standard errors of functions of random variables. One promising approach outlined in Kendall and Stuart (1969, p. 231) involves evaluating terms of a Taylor expansion. Using the Kendall and Stuart procedure it should be possible to derive expressions for the standard error of measurement for each of the models. This will allow for formal comparison of the models without real or simulated data.

The discussion then considered whether it was possible to compare the models by obtaining an estimate of "true score" and comparing it to the "real" true score. The problem lies in obtaining an acceptable true score. Three approaches were considered and are expected to provide a basis for future research. The first is to base model comparisons on Monte Carlo simulation studies. Monte Carlo studies provide an unambiguous true score but suffer from their lack of generalizability to practical applications. A second approach is to define true score as the score obtained on an instrument consisting of a large number of items. The models would then be used to estimate the true score using a smaller and more realistic number of items. This approach is empirical and more directly oriented to practical applications where testing time and the number of items that may be included in an instrument are limited. Although this approach suffers from the fact that the defined true score is not error free, the amount of error is not likely to be significant for practical purposes. The third approach would investigate the possibility of applying Geisser's predictive sample reuse method (Geisser, 1975) to the comparison of the models. Geisser's method may provide a more formal empirical approach to model comparison than the second approach discussed above, however, it has not been determined whether or not it is applicable to this research.

Four models for estimating true scores were presented and methods for comparing their outputs were discussed. Procedures for comparing the statistical properties of the models are available and relatively straightforward. Future research will be concerned with establishing the empirical validity of the models and their applicability to solving practical measurement problems.

REFERENCES

Epstein, K.I. An empirical investigation of four criterion-referenced testing models. Presented at 17th Annual Conference of the Military Testing Association, Indianapolis, IND, Sept., 1975.

Geisser, S. The predictive sample reuse method with applications. Journal of the American Statistical Association, 1975, 70, 320-328.

Kendall, M.G. & Stuart, A. The advanced theory of statistics: 3rd Edition (Vol. 1). New York: Hafner Publishing Company, 1969.

Lewis, C., Wang, M.M., and Novick, M.R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin 13. Iowa City, Iowa: American College Testing Program, 1973.

Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, MASS: Addison-Wesley, 1968.

Wright, B., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.